BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

22/6

09/868554

#4

842

PCT/JP99/07050

日本国特許庁

JP99/7050

EJU

PATENT OFFICE
JAPANESE GOVERNMENT

24.01.00

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: １９９８年１２月１５日

出願番号
Application Number: 平成１０年特許願第３５５６５７号

出願人
Applicant (s): 松下電器産業株式会社

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

２０００年　２月２５日

特許庁長官
Commissioner,
Patent Office 近藤隆彦

出証番号　出証特２０００－３００９４９６

【書類名】　特許願

【整理番号】　2015200176

【提出日】　平成10年12月15日

【あて先】　特許庁長官殿

【国際特許分類】　G06F 15/403

【発明の名称】　検索処理方法

【請求項の数】　14

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　今川　太郎

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　松川　善彦

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　近藤　堅司

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　目片　強司

【特許出願人】

【識別番号】　000005821

【氏名又は名称】　松下電器産業株式会社

【代理人】

【識別番号】　100097445

【弁理士】

【氏名又は名称】　岩橋　文雄

【選任した代理人】

【識別番号】　100103355

【弁理士】

【氏名又は名称】　坂口　智康

【選任した代理人】

【識別番号】　100109667

【弁理士】

【氏名又は名称】　内藤　浩樹

【手数料の表示】

【予納台帳番号】　011305

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9809938

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　検索処理方法

【特許請求の範囲】

【請求項１】　１つ以上の文字およびまたは１つ以上の文字片から成る単位を文字要素とし、文書データ群の中から、指定した文字列を構成する文字要素とあらかじめ定めた関係を満たす文字要素列データを検索することを特徴とする検索処理方法。

【請求項２】　文書データ群に文字要素同士の接続関係を複数通り保持しておき、前記文書データ群の中から指定した文字列とあらかじめ定めた関係を満たす文字要素列データを検索することを特徴とする検索処理方法。

【請求項３】　接続する文字要素の位置を複数保持しておくことを特徴とする請求項２記載の検索処理方法。

【請求項４】　接続する文字要素を複数保持しておくことを特徴とする請求項２記載の検索処理方法。

【請求項５】　指定した文字列を構成する文字要素と文書データ群を構成する文字要素との距離があらかじめ定めた基準以下のデータを検索することを特徴とする請求項１～４いずれかに記載の検索処理方法。

【請求項６】　指定した文字列を構成する文字要素とあらかじめ定めた関係を満たす文字要素列データを文字要素間の距離を表現したテーブルを用いて検索することを特徴とする請求項１～５いずれかに記載の検索処理方法。

【請求項７】　複数のテーブルを有し、テーブルを切り替えて検索を行うことを特徴とする請求項６記載の検索処理方法。

【請求項８】　指定した文字列を構成する文字要素をテーブル中の他の文字要素に置き換えて文書データ群から検索することを特徴とする請求項６または７に記載の検索処理方法。

【請求項９】　文書データ群中の文字要素をテーブル中の他の文字要素にあらかじめ置き換えたデータを文書データ群に付加しておくことを特徴とする請求項６～８いずれかに記載の検索処理方法。

【請求項１０】　検索結果に基づいて動作を決定することを特徴とする請求項

９いずれかに記載の検索処理方法。

【請求項１１】　指定した文字列を構成する文字要素を検出した位置または位置関係に基づいて動作を決定することを特徴とする請求項１０記載の検索処理方法。

【請求項１２】　文書データ群中における検索すべき文字要素列の有無または位置に基づいたコマンドを出力することを特徴とする実施例１０記載の検索処理方法。

【請求項１３】　指定した文字列と検索すべき文字要素列との関係を異なる手段で再度判断することを特徴とする請求項１～１２いずれかに記載の検索処理方法。

【請求項１４】　請求項１～１３のいづれか一つの請求項に記載の各手続きの全部または一部の手続きをコンピュータに実行実現するためのプログラムを格納したことを特徴とする情報記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は文字認識した文書を含むデータから指定した文字列に基づき、前記文字列に関連したデータを検索し、活用する技術に関するものである。

【０００２】

【従来の技術】

従来の文字認識した文書を含むデータから指定した文字列に基づいて関連したデータを検索する技術としては特開平７-152774号公報「文書検索方法および装置」がある。

【０００３】

従来の方法の例を図１６を用いて説明する。図１６は紙に記された文書と前記文書を文字認識した場合の結果を示している。通常文字認識においては、紙面に印字された文字のかすれや傾き、字体、文字サイズなどの影響で認識誤りを生じ得る。

【０００４】

図１６においてはオリジナルの文書における「本」という文字が「木」という字に誤って認識されている。

【０００５】

また、オリジナルの文書における「口」という文字が「区」という文字に誤認識されている。

【０００６】

ここで、「日本」という文字列を検索する場合を考える。このとき、（表１）に示すような誤認識文字の表を用いる。

【０００７】

【表１】

対象文字	誤認識文字
本	木、大、太、才
口	口、回、円、々

【０００８】

誤認識文字の表はあらかじめ、文字認識によって間違われやすい文字を並べた表である。

【０００９】

（表１）においては、「本」という文字は「 木、大、太、才 」に誤って認識されやすく、「口」という文字は「 口(記号の四角形)、回、円、々 」に間違われやすいということを示している。

【００１０】

「日本」を検索する場合、文字認識された文書より「日本」という文字列を検索すると同時に誤認識文字の表を用いて「日木」、「日大」、「日太」、「日才」という文字列を生成し、これらの文字列も「日本」同様に検索することで「日本」が誤認識された「日木」の部分を望ましく検索できるようになる。

【００１１】

【発明が解決しようとする課題】

しかしながら上記従来の手法ではあらかじめ、誤認識しやすい文字を用意しておくために、誤りの少ない文書データを検索する時には余分な文字候補を用いた余分な検索処理が行われ、また逆に誤りの多い文書データではあらかじめ用意した誤認識文字表に含まれる文字以外の誤認識には対応できない場合が発生するという課題を有していた。

【００１２】

例えば、図１６において「人口」という文字列を検索したい場合、誤認識文字の表を用いて「人口(記号の四角形)」、「人回」、「人円」、「人々」を同時に検索するが、誤認識文字の表に存在しない誤り（「口」を「区」に間違う）が発生した場合である「人区」（本来は「人口」）は検索が不可能であった。

【００１３】

また、一般の文書を文字認識したデータを検索する場合、文字認識時の文書レイアウトの判断誤り（縦書きと横書きの誤判断、改行後の次行への接続の判断誤り、段落から段落への接続の判断誤りなど）が起こり得るが、上記手法ではレイアウトの誤りに対しては対応できないという課題を有していた。

【００１４】

例えば、図１７のようなレイアウトの文書を文字認識する場合を考える。図１７において段落の正しい順番は、右上の段落、左上の段落、右下の段落、左下の段落である。しかしながら、文字認識の過程において、段落の順番を誤って判断し、右上の段落の次に右下の段落が接続すると判断する場合が起こり得る。ここで「日本の人口」という文字列を検索したい場合、誤認識表などを用いて個々の文字について望ましい検索が可能であっても、段落の接続が誤っている場合、図１８のように「・・・日本のする傾向・・・」という文書として扱われるため、「日本の人口」という文字列は検索できない。

【００１５】

そこで、本発明は上記従来の課題を鑑み、検索したい文字列を複数の要素（文字または文字片または文字列または文字片と文字との組合わせ）に分け、それぞれを独立に検索することで、複数行にわたる文字列の検索を行う際に、レイアウトの誤認識が含まれていても検索を可能とする。

【００１６】

また、文字要素同士の関係をテーブルとしてを保持しておくことで、許容できる誤りの度合いが可変かつ高速な検索を可能とする。

【００１７】

【課題を解決するための手段】

本発明は１つ以上の文字およびまたは１つ以上の文字片から成る単位を文字要素とし、文書データ群の中から、指定した文字列を構成する文字要素とあらかじめ定めた関係を満たす文字要素列を検索することを特徴とする検索処理方法である。

【００１８】

また、文字要素間の距離を表現したテーブルを用いて、指定した文字列を構成する文字要素とあらかじめ定めた関係を満たす文字要素列を検索すること検索処理方法である。

【００１９】

【発明の実施の形態】

（第１の実施の形態）

以下、本発明の実施の形態について図面を参照して説明する。図１は本発明の第１の実施の形態で用いる文字要素同士の距離を示す距離テーブルの一例を示すものである。

【００２０】

ここで、文字要素とは「亜」のような文字そのものや、「00」のような複数の文字の集まったもの、図２のような文字を構成する文字片や、図３のような文字片と文字とが集まったものなどを示す。

【００２１】

また、文字には「）」や「◎」のような記号も含める。

距離テーブルは文字要素間の近い・遠いの関係を数値で表現したものである。

【００２２】

図１では文字要素「亜」と文字要素「唖」との距離が１０、文字要素「亜」と文字要素「00」との距離が１７２であることを示しており、文字要素「亜」は文

字要素「00」よりも文字要素「唖」に距離が近いことを示している。他の文字要素についても同様に文字要素間の距離を定義しておく。

【００２３】

距離の定義に関しては、特定の文字認識システムの入出力関係、各文字要素の形状を特徴量数値で表現した場合の特徴量空間内でのユークリッド距離などを用いることができる。

【００２４】

また、距離テーブルは文字要素間の距離を表現していれば、必ずしも図１のような格子状の表の形態をとる必要はなく、文字要素ごとに他の文字要素との距離を距離の近い順番に保持していても良く、また順番そのものを距離と扱うことも可能である。

【００２５】

第一の実施の形態について、文字要素間距離テーブルを用いた文字列の検索につて、その手続きを説明する。

【００２６】

ここで、図７のように本来「・・・日本の人口構成は・・・」である文書を文字認識して得られた「・・・日木の人区構成は・・・」という認識結果（文書データ、あらかじめ記憶媒体などに保持しておく）から文字列を検索することを考える。

【００２７】

通常、文字認識技術を用いた場合様々な要因で誤りが生じる。この場合、「本」という字が誤って「木」に認識され、「口」という文字が誤って「区」という文字に誤って認識されている。

【００２８】

ここで、「日本」という文字列を指定して、本来「日本」が存在した場所を図７の認識結果の中から検索することを考える。

【００２９】

まず指定した文字列「日本」を構成する文字要素として「日」について文字要素間距離テーブルを参照し、距離があらかじめ定めた値（例えば１５０など）よ

りも小さい文字（例えば「日」については「日」と「目」）を図７の認識結果から検索する。

【００３０】

この場合、認識結果の中の「日」という文字要素を結果として検出する。次に指定した文字列「日本」を構成する次の文字要素として「本」について文字要素間距離テーブルを参照し、距離があらかじめ定めた値（同上の１５０）よりも小さい文字（「本」については例えば「本」と「木」と「大」）が「日」を検出した位置の次の位置の文字要素「木」と一致するかを図７の認識結果から判断する。

【００３１】

この場合、文字要素「木」が一致することから、指定した文字列「日本」に対して図７の認識結果中の「日木」が検出できる。

【００３２】

これによって本来「日本」という文字列が誤って「日木」と認識された場合にでも、本来の文字列の位置を検索することが可能となる。実際には指定した文字要素列を検出した場合、検出した文字要素列のみならず、検出した文字要素列を含む前後の文書の認識結果も合せて検索者に提示したり、文字認識を行ったオリジナルの文書を画像イメージとして別途文書データに保持しておき、対応する画像イメージを検索者に提示することで、例え文字認識の結果が一部誤っていた場合にでも、人間が再度判断することで、検索者が必要とする情報を得ることが可能となる。

【００３３】

また、指定した文字要素列を検出した際には、前後の文書の提示以外に文書のタイトルや要約を表示しても良い。この場合、少ない表示スペースで検索結果を把握することが可能となる。また、表示以外に音声を用いて前後の文章やタイトル、要約を出力することで、表示領域の少ない端末にも対応することができる。

【００３４】

また、出力は通信路（ネットワーク）を経由して出力しても良い。また、帯域が狭い通信路を経由する場合には、検索結果の画像イメージを最初から表示する

のではなく、前後の文書の認識結果やタイトル・要約のみを最初に表示し、検索者が別途指示することで情報量の多い画像イメージの表示を行うことで、検索時間や閲覧時間を節約することが可能となる。

【００３５】

更に、指定した文字要素列が検出できた際に、検出した情報を提示するのではなく、機器への新たな命令（コマンド）を発行しても良い。例えば、リアルタイムにカメラなどから得られる画像に対して検索を行い、特定の文字要素列を検出した場合（例えば「レストラン」）に撮像を行う機器に対して映像をメモリに記録するコマンドを発行したり（レストランの映像を集めることが可能となる）、特定の文字要素列を検索した場合には、文字要素列を含む画像をプリンターへ印刷するコマンドをプリンタへ発行することや、文字要素列を含む画像の情報を通信網（ネットワーク）を通して複数の宛名に配信することなどが考えられる。

【００３６】

なお上記の場合、文字要素間距離について１５０という値をあらかじめ定めたが、この値は可変であり、最初に小さい値を設定して検索を行い、検索できない場合に順次大きな値に再設定して検索しても良い。これは最初に小さい距離に相当する値を設定することで、文字認識について誤りをあまり許容しない状態で検索を行い、順次文字認識の誤りを許容して検索することに相当する。よって、最初から文字認識の誤りを大きく許容することで、関係の無い余分な文字要素列の検出が発生することを未然に防ぐことが可能となる。

【００３７】

また、認識結果を保持するデータ（文書データ）に認識の信頼度も合せて保持しておくことで、前記信頼度に応じて検索に用いる距離の基準値を適切な値に設定することが可能である。例えば、認識の信頼度の低い文書については、検索時の文字要素間距離テーブルで許容する距離を大き目に設定し、逆に認識結果の信頼度が高い場合には、許容する距離を小さ目に設定することで、関係の無い余分な文字列要素の検出を抑えることが可能となる。信頼度の設定については文書毎や文字毎に付与することも出来る。また、文字認識の信頼度は、文字認識を行う際の認識システム（例えばニューラルネットワークなど）の出力などを用いるこ

とが可能である。

【００３８】

ここでは、指定した文字列を構成する文字要素の先頭「日」から順番に一文字ずつ検索したが、異なる順番でも良い。特に一般的な文書中に出現する頻度を考慮し、指定した文字列を構成する文字要素のうち、一般的に文書中に出現する頻度の低い文字要素から検索することで、余分な検索手続きを減らすことが可能となり、検索速度を速めることが可能となる。

【００３９】

なお、上記の例では文書データを記憶媒体（メモリや磁気ディスク、光ディスクなど）にあらかじめ保持してくことを想定したが、画像入力機器（スキャナ、デジタルカメラ、ビデオカメラなど）から入力した画像情報を逐次文字認識して得られるリアルタイムの情報について同様の検索を行っても良い。

【００４０】

このように、文字要素間の距離テーブルを用いて文字要素列の検索を行うことで、指定した文字要素列が誤認識で他の文字要素に置き換わったような文字要素列を文書データから検索することが可能となる。距離テーブルを用いることで、複雑な距離計算などを行う必要もなく高速な検索が可能となる。また、距離テーブルを用いることで、誤認識の許容度合いを都度適切な値に設定することが出来、効率の良い検索が可能となる。

【００４１】

（第２の実施の形態）

次に、本発明の第２の実施の形態について説明する。

【００４２】

第２の実施の形態は複数の文字要素間距離テーブルを用いた例である。

検索の基本的手続きは第一の実施の形態と同様である。

【００４３】

第２の実施の形態においては、複数の異なる文字要素間距離テーブルを用いて検索を行う。

【００４４】

複数の文字要素間距離テーブルとしては、複数種類の文字認識システムそれぞれに対応したテーブルや、文字種（漢字、アルファベット、ギリシャ文字、カタカナなど）ごとに対応したテーブル、フォント種（ゴシック体のテーブル[図１]、明朝体のテーブル[図８]など）ごとに対応したテーブルなどをあらかじめ用意し、検索する文書データに応じて用いるテーブルを切り替える。例えば、文字認識して得た文書データについてあらかじめ、そこに含まれる文字の字種やフォントの種類、用いた認識システムの種類などを文書データに別途保持しておくことで、検索の際に適切なテーブルを選択して用いることが可能となり、検索の精度と速度を向上させることが可能となる。

【００４５】

フォントごとにテーブルを切り替える場合、文字認識して得た文書データにあらかじめ文字ごとに明朝体に近いかゴシック体に近いかの情報を付与しておき、ゴシック体の文字から成る文書データより文字要素列を検索する場合には図１のようなテーブルを用い、明朝体の文字から成る文書データより文字要素列を検索する場合には図８のようなテーブルを用いる。フォントの種類の情報については文字認識を行う際に同時にフォントの種類を認識することなどにより得ることができる。

【００４６】

また、同一の文書データに対しても複数のテーブルを切り替えても良い。この場合、特定のテーブルを用いて検索できなかった文書データについて再度、検索したい文字要素列の有無を異なる尺度で検証し、検索の精度を向上させることが可能となる。更に、テーブルを用いた検索の結果に対して、前記検索で得られた文字要素列の位置に対応する文書の画像イメージを用いて再度高精度な文字認識を行っても良い。これにより、テーブルを用いて高速な粗検索を行い候補を絞った後に、高精度な文字認識（一般的に処理時間がかかる）を用いて検索対象を確定することが可能となり、検索精度と検索速度の両立が可能となる。

【００４７】

なお、上記例では１文字単位の文字認識の誤りを補う為に文字要素間距離テーブルを用いていたが、文字認識を行う際には複数の文字要素を単一の文字として

扱った結果生じる誤認識（図４：２つの「木（き）」を「林（はやし）」と認識、２つの「０（ゼロ）」を「∞（無限大）」と認識）や、逆に単一の文字を複数の文字として扱った結果生じる誤認識（図５：「川」を３つの「１（いち）」と認識、「い」を「し」と「１（いち）」と認識）が存在する。このような場合にも、文字要素距離テーブルに「木（き）」２文字と「林」との距離、「０（ゼロ）」２文字と「∞（無限大）」との距離、「川」と３つの「１（いち）」、「い」と「し」、「１（いち）」との距離がそれぞれ小さいことを保持しておく。

【００４８】

図６に図４や図５の例を含む文字要素間距離テーブルの一例を示す。図６では「木（き）」２文字と「林」との距離、「０（ゼロ）」２文字と「∞（無限大）」との距離、「川」と３つの「１（いち）」、「い」と「し」、「１（いち）」との距離はそれぞれ１３以下で、他の組み合わせの場合（距離９８以上）よりも小さい値になっている。

【００４９】

ここで、検索したい文字要素列として「１００」を指定した時に、誤りの度合いの基準となる距離を５０とすると、「１」を検索した後、「０」を２つ検索すると同時に「∞」も検索することで、「１００」が「１∞」と誤認識された文字要素列を検索することが可能となる。同様に、検索したい文字要素列として「いろり」を指定した時に、誤認識された「し１ろり」という文字要素列を文書データから検出することなども可能となる。

【００５０】

更に、かな漢字変換等の誤りによってオリジナルの文書自体に文章として誤った表現が含まれる場合（「納める」を「収める」と表現）や、複数の送り仮名付け方が存在する場合（「変る」を「変わる」）や、漢字表記した言葉をひらがなで検索しようとする場合（「切磋」を「せっさ」で検索）や、類義語で検索しようとした場合（「価格」を「定価」で検索）や、異なる言語に対して検索しようとした場合（「history」を「歴史」で検索する場合）についても、文字要素間距離テーブルにおいてそれぞれ「収」と「納」との距離、「変わ」と「変」との距離、「切磋」と「せっさ」との距離、「価格」と「定価」のと距離、「histor

y」と「歴史」との距離を小さな値として定義しておくことで検索することが可能となる。

【００５１】

（第３の実施の形態）

次に、本発明の第３の実施の形態について説明する。

【００５２】

第３の実施の形態は文字要素間距離テーブルを用いて指定した文字要素列をあらかじめ複数の文字要素列に置き換えて検索を行う例である。

【００５３】

図７の認識結果（文書データ）から指定した文字要素列として「日本」を検索する場合を考える。最初に文字要素列「日本」を構成する文字要素「日」と「本」に分け、それぞれの文字要素について文字要素間距離テーブルを参照し、距離があらかじめ定めた値（例えば１５０など）よりも小さい文字（例えば「日」については「日」と「目」、「本」については例えば「本」と「木」と「大」）同士を組み合せて新たな文字要素列「日本」、「目本」、「日木」、「目木」、「日大」、「目大」を生成する。

【００５４】

次に前記生成した文字要素列それぞれについて図７の認識結果（文書データ）より検索を行う。この場合、生成した「日木」がオリジナルの文章で「日本」が存在する位置において検出でき、望ましい検索結果を得ることが出来る。

【００５５】

ここで、検索した結果何も検出されない場合には、再度距離の基準値を大きくし（例えば２００に設定する）、新たにより多くの文字要素列を生成して同様な検索することで、距離テーブルで許容する距離が１５０の時には検出できなかった認識誤りを検出することも可能である。

【００５６】

このように、文字要素間距離テーブルを用いて指定した文字要素列を誤りの可能性のある複数の文字要素列に置き換えて検索することでも第一の実施の形態と同様に指定した文字要素列が誤認識で他の文字要素に置き換わったような文字要

素列を文書データから検索することが可能となる。距離テーブルを用いることで、複雑な距離計算などを逐次行う必要がなく高速な検索が可能となる。また、距離テーブルを用いることで、誤認識の許容度合いを都度適切な値に設定することが出来、効率の良い検索が可能となる。

【００５７】

（第４の実施の形態）

次に、本発明の第四の実施の形態について説明する。

【００５８】

第４の実施の形態は文字要素間距離テーブルを用いて認識結果の文書データ中の文字要素に他の複数の文字要素を付加した後検索を行う例である。図７の認識結果（文書データ）から指定した文字要素列を検索する場合を考える。

【００５９】

ここで、あらかじめ文字要素間距離テーブルを参照し、認識結果の文書データの各文字要素について距離があらかじめ定めた値（例えば１５０など）よりも小さい文字（例えば「日」については「日」と「目」、「木」については「本」と「大」など）を付加しておく（図９）。

【００６０】

次に指定した文字要素列として「日本」を検索する場合には、「日本」を「日」と「本」に分け、最初に「日」を検索する。この場合図９の文書データの１列目で「日」が検出できる。次に「日」を検出した位置の次の文字要素（「木」、「本」、「大」）の中に「本」が存在するかを判断し、「本」が含まれているので「日本」という文字要素列が検出できたとする。文字要素列の要素数が３つ以上の時にも同様な手続きで検出を行い、全ての文字要素が連続して検出できた場合に指定した文字要素列を検出できたと判断する。

【００６１】

このように、本実施の形態においては文字要素間距離テーブルを用いて認識結果の文書データ中の文字要素にあらかじめ複数の文字要素を付加しておくことで第一の実施の形態と同様に指定した文字要素列が誤認識で他の文字要素に置き換わったような文字要素列を文書データから検索することが可能となる。また、あ

らかじめ文書データ中の各文字要素に複数の異なる文字要素を付加しておくことで、検索時に距離テーブルを参照する手続きを省くことが可能となる。

【００６２】

なお、上記の形態に加えて、第一や第二の実施の形態のように検索の過程で距離テーブルを用いる形態や、第三の実施の形態のように指定した文字要素列を距離テーブルを用いてあらかじめ他の文字要素列に置き換える形態を併用して実施することも可能である。

【００６３】

（第５の実施の形態）

次に、本発明の第５の実施の形態について説明する。

【００６４】

図１０のようなレイアウトの文章を文字認識した結果（文書データ）から文字要素列「日本の人口」を検索する場合を考える。

【００６５】

図１０において文章の正しい順序は段落Ａ、段落Ｂ、段落Ｃ、段落Ｄの順番である。（一般的に様々なレイアウトを想定した場合、自動的に段落同士の接続の正しい順番を決めることは難しく、接続の誤りが発生し得る。）ここで、文字認識した結果を各段落ごとに付与した固有の番号（図１０ではＡ，Ｂ，Ｃ，Ｄ）と共に格納しておく（図１１）。

【００６６】

このとき各段落について次に接続する可能性のある段落の番号も合せて記録しておく。図１１の場合、段落Ａの後に接続する可能性のある段落はＢとＣであることを意味している。

【００６７】

ここで、接続の可能性のある段落の決め方としては、オリジナルの画像を文字認識する際に各段落同士のそれぞれの位置関係を参照することで決める。

【００６８】

例えば、縦書きの場合、ある段落Ｘの次に続く段落としては段落Ｘよりも下に位置するか、左に位置する段落を接続の可能性のある段落とする。更に、文字認

識した結果に基づいて、ある段落Ｘの末尾の文章と文法的に接続可能な文頭を有する段落を接続の可能性のある段落としてもよい。また、レイアウトに特定の規則がある文章については、その規則に基づいた接続の可能性のある段落を選択することでもよい。

【００６９】

図１１のような認識結果（文書データ）から「日本の人口」という文字要素列を検索する場合、第一から第四までの実施の形態と同様に、「日本の人口」を構成する各文字要素「日」、「本」、「の」、「人」、「口」を各段落から順次検索していく。図１１の場合段落Ａの末尾に文字要素「日」、「本」、「の」が続けて検出できる。次に「人」を検索する。ここで、段落Ａに接続する可能性のある段落はＢとＣであるため、段落Ｂと段落Ｃのそれぞれの文頭に「人」が存在するかを判断する。この場合段落Ｂの文頭に「人」が検出できるので、その次の位置に「口」が存在するかを判断する。最終的に「日本の人口」を構成する全ての文字要素が検出できたことになる。

【００７０】

このように接続する可能性のある段落を複数保持しておくことで、文書認識の際に段落の接続を誤判断した場合にでも、複数の段落にまたがる文字要素列を検出することが可能となる。

【００７１】

また、段落同士の接続に限らず、段落内で行と行との接続があいまいな場合（行間に図、表、見出しなどが挿入される場合）にも同様に行ごとに異なる番号を付与し、接続する可能性のある行の番号を行ごとに複数保持しておくことで、複数の行にまたがる文字要素列を正しく検出することが可能となる。

【００７２】

また、文字要素と文字要素との接続があいまいな場合（図、表の挿入が間にある場合や、文字要素列の配置が装飾的な場合[曲線状に配置された文字要素列]など）にも同様に行（文字要素）ごとに異なる番号を付与し、接続する可能性のある行（文字要素）の番号を行（文字要素）ごとに複数保持しておくことで、複数の行（文字要素）にまたがる文字要素列を正しく検出することが可能となる。

【００７３】

なお、各段落（行または文字要素）についてその段落（行または文字要素）の前に接続する段落（行または文字要素）の番号を合せて保持する形態でも同様の効果が得られる。また、接続する段落（行または文字要素）の番号の表現としては上記のように段落（行または文字要素）番号の絶対値で表現する以外に、段落（行または文字要素）番号の相対値（段落Ａに接続する段落を段落Ｂ，段落Ｃと表現する代わりに、段落＋１、段落＋２と表現する）で表現しても良い。

【００７４】

（第６の実施の形態）

次に、本発明の第６の実施の形態について説明する。

【００７５】

第５の実施の形態と同様に、図１０のようなレイアウトの文章を文字認識した結果（文書データ）から文字要素列「日本の人口」を検索する場合を考える。

【００７６】

ここで、文字認識した結果を各段落ごとに付与した固有の番号（図１０ではＡ，Ｂ，Ｃ，Ｄ）および各段落の文書内での位置座標（図１０では右上を原点とし左方向をＸ座標、下方向をＹ座標とする）と共に格納しておく（図１２）。図１２では、段落Ａの位置座標（Ｘ、Ｙ）が（１０、１００）であることを示している。

【００７７】

検索する手続きは第五の実施の形態とほぼ同様であるが、段落Ａの末尾に文字要素「日」、「本」、「の」が続けて検出できた後に「人」を検索する際、段落Ａに接続する可能性のある段落を図１２に格納している段落の座標値を用いて決定する。この場合、段落Ａの座標値は（Ｘ、Ｙ）＝（１０、１００）であるので、隣接する段落の座標としてＸ座標値が等しく、Ｙ座標値が次に大きい段落Ｃ（Ｘ、Ｙ）＝（１０、２００）とＹ座標値が等しくてＸ座標値が次に大きい段落Ｂ（Ｘ、Ｙ）＝（１００、１００）を接続する可能性のある段落と判断し、段落Ｂと段落Ｃのそれぞれの文頭に「人」が存在するかを判断する。

【００７８】

この場合段落Ｂの文頭に「人」が検出できるので、その次の位置に「口」が存在するかを判断する。最終的に「日本の人口」を構成する全ての文字要素が検出できたことになる。

【００７９】

接続する段落を決定する方法としては、上記の例以外にも、縦書きの場合、ある段落Ｘの次に続く段落としては段落Ｘよりも下に位置するか、左に位置する段落を接続の可能性のある段落としても良い。

【００８０】

また、レイアウトについて特定の規則がある文章については、その規則に基づいた接続の可能性のある段落を位置座標を用いて選択する。

【００８１】

なお、位置座標の表現としては、原点や座標軸は自由に選んで良く、また座標値についても段落や図ごとに番号を割り振った値の順番を座標値の単位として用いてもよい。

【００８２】

このように段落ごとに段落の位置の情報を保持しておくことで、文書認識の際に段落の接続を誤判断した場合にでも、接続する可能性のある段落を選択し、複数の段落にまたがる文字要素列を検出することが可能となる。また、位置座標を保持しておくことで、文書データを変更することなく接続する段落の決定方法を変更することが可能であり、段落の位置座標は文書のレイアウトを再現する為に用いることも可能である。

【００８３】

なお、上記の例では段落ごとに位置座標を保持したが、行単位や文字要素単位に異なる番号を付与し、位置座標と共に格納して、接続する可能性のある行または文字要素を決定して検索することで、複数の行または文字要素にまたがる文字要素列を検索することが可能となる。

【００８４】

（第７の実施の形態）

次に、本発明の第七の実施の形態について説明する。

【００８５】

第５の実施の形態と同様に、図１０のようなレイアウトの文章を文字認識した結果（文書データ）から文字要素列「日本の人口」を検索する場合を考える。

【００８６】

ここで、文字認識した結果は図１３のように接続する可能性のある段落との可能性を含めた複数の認識結果として文字認識結果（文書データ）に保持しておく。

【００８７】

図１３では文字認識結果として２種類保持しており、段落Ａに段落Ｂが接続した場合（文字認識結果２）と段落Ａに段落Ｃが接続した場合（文字認識結果１）とを保持している。

【００８８】

図１３の中から「日本の人口」を検索する場合、文字認識結果１と２それぞれに対して「日本の人口」を検索し、文字認識結果２から「日本の人口」を検出することができる。

【００８９】

なお、図１３のように複数の段落の接続を想定して認識結果を保持する場合、検索する文字要素数に上限を設け（例えば１０文字要素）、段落Ａに接続する段落Ｂ、段落Ｃの認識結果は段落の文頭から９文字要素のみを段落Ａの認識結果に付加して保持するようにしてもよい。この場合段落Ａから段落Ｂまたは段落Ｃにまたがる１０文字要素までの文字要素列の検索が可能となる。

【００９０】

このように、あらかじめ接続する可能性のある段落を含めて認識結果を複数保持しておくことで、文書認識の際に段落の接続を誤判断する場合にでも、複数の段落にまたがる文字要素列を検出することが可能となる。また、段落間の接続を含めて複数の認識結果を文書データに保持しておくことで、検索手続きは簡易になり、従来法の検索手続きを利用することが可能となる。

【００９１】

（第８の実施の形態）

次に、本発明の第８の実施の形態について説明する。

【００９２】

図１４のようなレイアウトの文章を文字認識した結果（文書データ）から文字要素列「神戸」を検索する場合を考える。

【００９３】

図１４のような文書の場合、本来の文書の向きは横書きであるが、文字同士の間隔は縦方向に接近しており、文字認識を行った場合に誤判断する可能性がある。

【００９４】

ここで、各文字要素を文字認識した場合に、縦書きを想定した場合と横書きを想定した場合の２種類のレイアウトに対応した認識結果（図１５のａ、ｂ）を作成し、認識結果の文書データとして保持しておく。ここで、「神戸」という文字要素列を縦書き、横書きそれぞれに対応した認識結果から検索を行い、横書きを想定した図１５（ａ）の３行目から「神戸」を検出し、図１４の文書に「神戸」が含まれることが分かる。

【００９５】

このように、複数のレイアウトに対応した認識結果を文書データに保持しておくことで、レイアウトの判断が困難な文書に対しても認識結果に基づいた文字要素列の検索が可能となる。また、検索手続きは従来法の検索手続きを利用することが可能となる。

【００９６】

なお、上記の例では縦書きと横書きの２種類のレイアウトを想定したが、縦書き・横書き以外にも斜め方向のレイアウトなどその他のレイアウトも同様に扱うことが可能である。

【００９７】

（第９の実施の形態）

次に、本発明の第９の実施の形態について説明する。第九の実施の形態ではレイアウトの判断を誤った認識結果の文書データから指定した文字要素列を検出する手続きを説明する。

【００９８】

図１４のようなレイアウトの文章を文字認識した結果（文書データ）から文字要素列「神戸」を検索する場合を考える。ここで、縦書きを想定した認識結果図１５（ａ）を認識結果（文書データ）として保持していたとする。

【００９９】

ここで、「神戸」という文字要素列を構成する文字要素「神」と「戸」とに分けて、それぞれの文字要素について図１５（ａ）から検索を行う。各文字要素の検索を行うと「神」は行番号５の第３文字目に検出でき、「戸」は行番号４の第３文字目に検出できる。ここで、構成する文字要素「神」、「戸」が全て検出できた場合、それぞれの文字要素を検出した位置関係に基づいて、文字要素列「神戸」の検出を判断する。

【０１００】

この場合「神」と「戸」が隣接する行の同じ文字数目に連続して検出できたため、文字要素列「神戸」が検出できたとする。個々の文字要素を検出した位置関係としては上記以外の基準を設けても良い。例えば、文字の位置座標が分かっている場合には、個々の文字要素があらかじめ定めた距離以下で接近しかつ直線的に配置してあることを判断基準としても良い。また、検索の手続きとしては他の手続きでも良く、上記のように文字要素を全て検索するのではなく、「神」を検出できた場合にのみ「神」を検出した行に隣接する行からのみ「戸」を検索するようにしても良い。これにより不用な検索手続きを削減し、効率的な文字要素列の検索が可能となる。

【０１０１】

このように、検索したい文字要素列を構成する文字要素を個別に文書データから検索し、個々の検出位置の位置関係から文字要素列の有無を判断することで、レイアウトの判断を誤って認識した文書データからでも指定した文字要素列を正しく検索することが可能となる。

【０１０２】

（第１０の実施の形態）

次に、本発明の第１０の実施の形態について説明する。

【０１０３】

第十の実施の形態では段落同士の位置関係を認識結果と共に文書データに保持しておき指定した文字要素列を検出する手続きを説明する。

【０１０４】

ここで、図１０のようなレイアウトの文章を文字認識した結果（文書データ）から文字要素列「日本の人口」を検索する場合を考える。第六の実施例と同様に、文字認識した結果を各段落ごとに付与した固有の番号（図１０ではＡ，Ｂ，Ｃ，Ｄ）および各段落の文書内での位置座標（図１０では右上を原点とし左方向をＸ座標、下方向をＹ座標とする）と共に格納しておく（図１２）。

【０１０５】

最初に、指定した文字要素列「日本の人口」を途中で分割し、２つの文字要素列に分ける（例えば「日本」と「の人口」など）。次に分割してできた２つの文字要素列それぞれを個々の段落から検索する。全ての分割の仕方について同様の手続きで検索を行う。「日本の」と「人口」に分割した場合、図１０では段落Ａの末尾に「日本の」が検出でき、段落Ｂの文頭に「人口」が検出できる。分割した文字要素列が全て検出できた場合、検出できた段落同士の位置関係に基づいて、文字要素列「日本の人口」の検出を判断する。例えば、２つの文字要素列を検出した段落が隣接していたり、位置が近い場合には指定した文字要素列を検出できたとする。図１０の場合、「日本の」が検出できた段落Ａの位置座標（ｘ、ｙ）＝（１０，１００）と「人口」が検出できた段落Ｂの位置座標（ｘ、ｙ）＝（１００，１００）は同じｙ座標で隣接することから「日本の人口」が検出できたとする。また、上記のように文字要素列を分割して得た文字要素列を段落内から検索する場合には、各段落の文末と文頭のみに検索処理を行えば検索効率が良い。

【０１０６】

なお、上記の例は指定した文字要素列を２つに分割したが必要に応じて３つ以上に分割しても同様の手続きが可能である。

【０１０７】

また、上記の例では複数の段落にまたがる文字要素列を検索する例を示したが

、複数の行にまたがる文字要素列についても同様に検索が可能である。この場合指定した文字要素列を分割し、各行に対して分割した文字要素列それぞれを検索し、分割した全ての文字要素列が隣接して検出できた場合にもとの指定した文字要素列が検出できたとする。

【０１０８】

このように本実施例では、段落や行の接続が誤っている（または不定な）場合にでも、複数の段落にまたがる文字要素列を正しく検出することが可能となる。

【０１０９】

このように本発明では、文字認識の誤りがある場合や、段落間・行間の接続が誤っていたり不定な場合や、縦書き・横書きの判断が誤っているあるいは不定である場合に指定した文字要素列を検索することが可能である。

【０１１０】

なお、本発明の第一から第十の実施の形態は単独で用いてもよいし、組み合わせて実施することも可能である。また、上記実施の形態の実現手段としてはハードウェアを用いて実現してもよいし、あるいはコンピュータ上のソフトウェアを用いて実現してもよい。

【０１１１】

また、上記実施の形態の各手続きのうち、全部または一部の手続きをコンピュータに実行させるためのプログラムを記録した媒体を用いる、あるいは通信網（ネットワーク）または放送を通じてプログラム（又はその一部）をダウンロードして実行することでも、上記の場合と同様の効果を実現することが可能である。

【０１１２】

【発明の効果】

以上のように本発明は、文字要素（文字または文字片または文字列または文字片と文字との組合わせ）同士の距離を表現したテーブルを用いることで、認識結果に対して許容できる誤り度合いを動的に変更して検索を行うことが可能である。

【０１１３】

また、テーブルを用いることで、複雑な距離計算を行わず高速な検索が可能である。

【０１１４】

また、文字要素同士の接続関係を複数保持する、あるいは複数通り検索する、あるいは文字要素列を分割して検索することで、文書のレイアウトを誤って解釈した文書データから望ましい検索を実現することが可能である。これにより、縦

書き・横書きを間違って判断している文書データや、改行後に継続する行を誤って判断している文書データからの文字列検索が可能となる。

【図面の簡単な説明】

【図１】

本発明の第一から第四の実施の形態で用いる文字要素間距離テーブルの一例を示す図

【図２】

文字片の例を示す図

【図３】

文字片と文字とが集まった文字要素の例を示す図

【図４】

複数の文字要素を単一の文字として扱った結果生じる誤認識の例を示す図

【図５】

単一の文字を複数の文字として扱った結果生じる誤認識の例を示す図

【図６】

第二の実施の形態で用いる文字要素間距離テーブルの一例を示す図

【図７】

検索する文書の例と認識結果の一例を示す図

【図８】

第二の実施の形態で用いる明朝体文字の文字要素間距離テーブルの一例を示す図

【図９】

第四の実施の形態で用いる複数の文字候補を保持した文書データの一例を示す図

【図１０】

検索する文書の例を示す図

【図１１】

第５の実施の形態で用いる認識結果を保持した文書データの一例を示す図

【図１２】

第６、第１０の実施の形態で用いる認識結果を保持した文書データの一例を示す図

【図１３】

第７の実施の形態で用いる認識結果を保持した文書データの一例を示す図

【図１４】

検索する文書の例を示す図

【図１５】

第８、９の実施の形態で用いる文書データの一例を示す図

（ａ）認識結果を縦書きとして保持した文書データを示す図

（ｂ）認識結果を横書きとして保持した文書データを示す図

【図１６】

検索する文書の例と認識結果の例を示す図

【図１７】

検索する文書の例を示す図

【図１８】

検索する文書の例と認識結果の例を示す図

【書類名】　図面

【図１】

	亜	唖	彡	卩	００
亜		１０	１３２	１６６	１７２
唖			１１５	１５２	１６４
彡				１４３	１９１
卩					６９
００					

【図２】

「彡」、「卩」

【図３】

「）彡」、「卩１」

【図４】

「木」、「木」→「林」
「０」、「０」→「∞」

【図５】

「川」→「１」、「１」、「１」
「い」→「し」、「１」

【図６】

	林	∞	し１	１１１	川
木木	１０	２２１	１９０	１５６	１５２
川	１５５	１６５	９１	９	
い	２０１	１１９	１３	８９	９５
卩	１４９	１８８	９８	１３３	１３７
００	２１５	１２	１０５	１６９	１７２

【図７】

オリジナル文書	・・・日本の人口構成は・・・
文字認識結果	・・・日本の人区構成は・・・

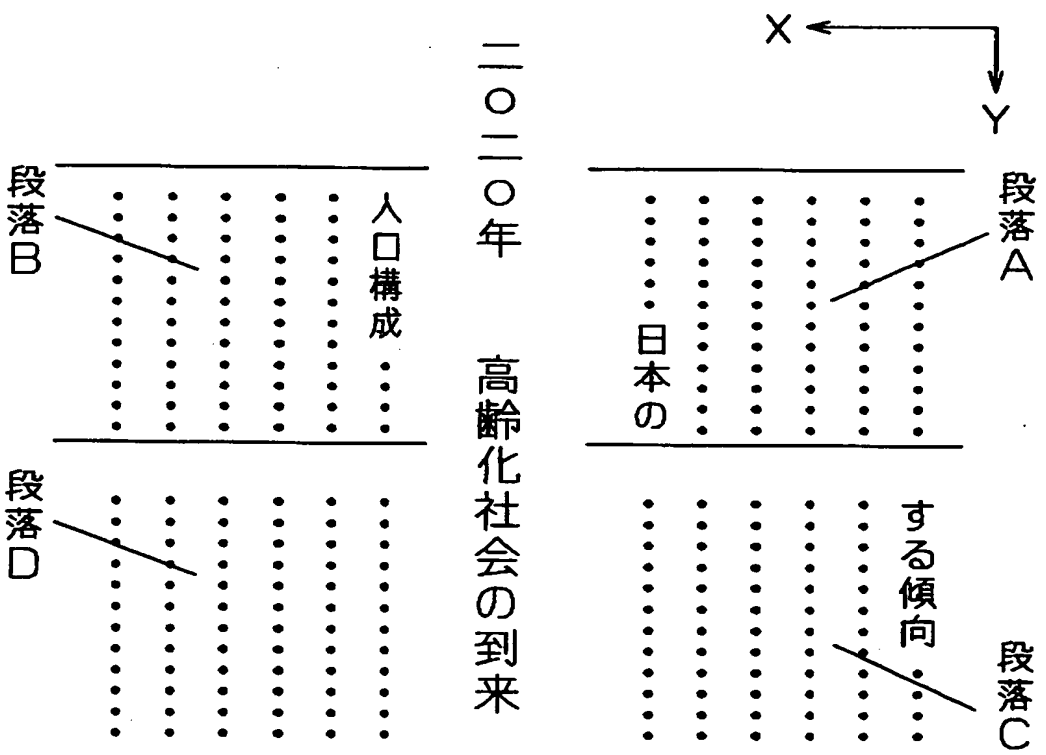
【図８】

	亜	唖	彡	卩	００
亜		１２	１３０	１７０	１６８
唖			１１４	１５０	１７０
彡				１４７	１９０
卩					６０
００					

【図９】

認識結果	・・・	日	木	の	人	区	構	成	は	・・・
候補		目	本	@	入	凶	講	茂	ほ	
候補			大		ル	図		感	ぼ	
候補						口				

【図１０】



【図１１】

段落番号	接続する段落の番号	段落単位の認識結果
A	B.C	・・・・・日本の
B	C.D	人口構成は ・・・
C	D	する傾向 ・・・・
D		・・・・・・・・・

【図１２】

段落番号	段落単位の認識結果	段落の位置	
		x	y
A	・・・・・日本の	１０	１００
B	人口構成は ・・・	１００	１００
C	する傾向 ・・・・	１０	２００
D	・・・・・・・・・	１００	２００

【図１３】

文字認識結果１	・・・日本のする傾向 ・・・
２	・・・日本の人口構成は・・・

【図１４】

京　都　　２９℃
大　阪　　３２℃
神　戸　　３０℃

【図１５】

(a)

列番号	列単位の認識結果
1	℃℃℃
2	920
3	233
4	都阪戸
5	京大神

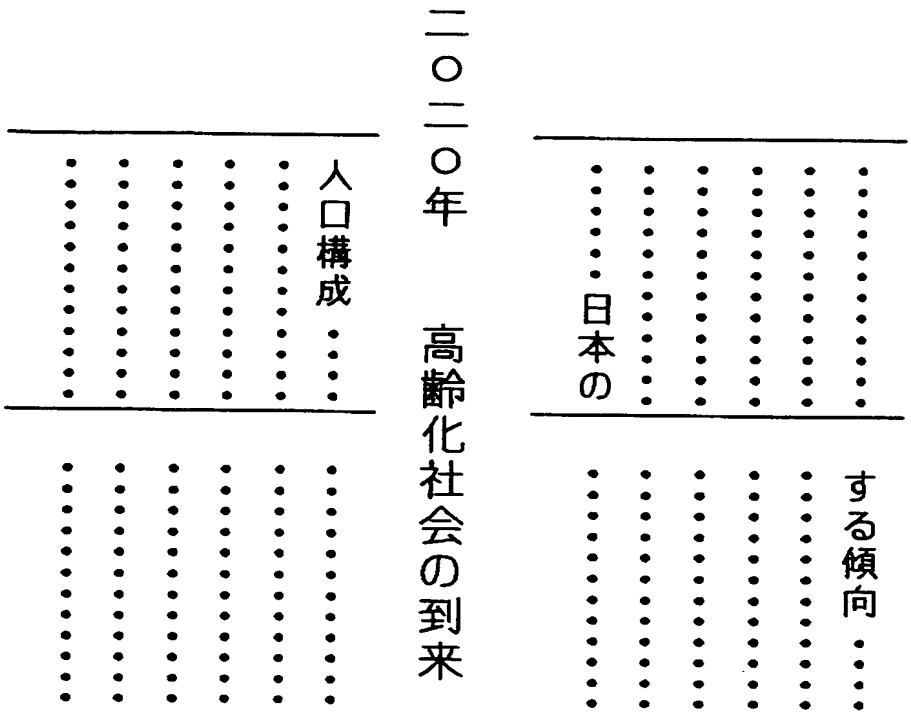
(b)

行番号	行単位の認識結果
1	京都２９℃
2	大阪３２℃
3	神戸３０℃

【図１６】

オリジナル文書	・・・日本の人口構成は・・・
文字認識結果	・・・日本の人区構成は・・・

【図１７】



【図１８】

オリジナル文書	・・・日本の人口構成は・・・
文字認識結果	・・・日本のする傾向　・・・

【書類名】　要約書

【要約】

【課題】　文章を文字認識した場合に生じる誤りを含んだ認識結果に対して、指定した文字列を誤認識のない状態で検索する場合と同様に検出すること。

【解決手段】　１つ以上の文字およびまたは１つ以上の文字片から成る単位を文字要素とし、文書データ群の中から、指定した文字列を構成する文字要素とあらかじめ定めた関係を満たす文字要素列を検索する検索処理方法である。また、文字要素間の距離を表現したテーブルを用いて、指定した文字列を構成する文字要素とあらかじめ定めた関係を満たす文字要素列を検索する検索処理方法である。

【選択図】　図１

【書類名】　職権訂正データ
【訂正書類】　特許願

＜認定情報・付加情報＞
【特許出願人】
【識別番号】　000005821
【住所又は居所】　大阪府門真市大字門真１００６番地
【氏名又は名称】　松下電器産業株式会社
【代理人】　申請人
【識別番号】　100097445
【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社　知的財産権センター
【氏名又は名称】　岩橋　文雄
【選任した代理人】
【識別番号】　100103355
【住所又は居所】　大阪府門真市大字門真１００６番地松下電器産業株式会社内
【氏名又は名称】　坂口　智康
【選任した代理人】
【識別番号】　100109667
【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内
【氏名又は名称】　内藤　浩樹

出　願　人　履　歴　情　報

識別番号　　〔０００００５８２１〕

１．変更年月日　　１９９０年　８月２８日
［変更理由］　新規登録
住　所　　大阪府門真市大字門真１００６番地
氏　名　　松下電器産業株式会社